3D FFTs for electronic structure calculations: Mixed programming models and communication strategies for many core architectures

Andrew Canning*, J. Shalf, L-W. Wang, H. Wasserman, M. Gajbe, N. Wright and S. Anderson (COE, Cray)

CRD, NERSC & UC Davis*

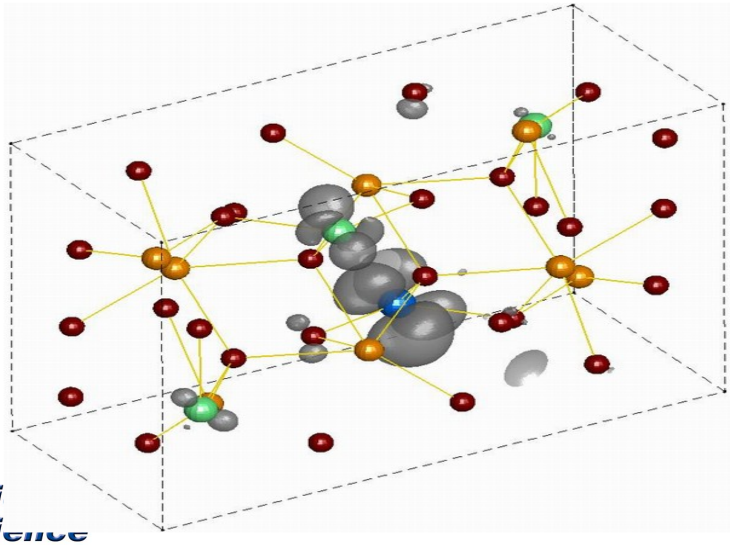
Office of Science
U.S. DEPARTMENT OF ENERGY

Outline

- Introduction to DFT Plane Wave Electronic Structure Calculations
- Parallel Data layouts and communication structures
- Scaling of our 3d FFT on various computers (Cray XT, IBM BG)
- Mixed OpenMP/MPI vs. MPI
- Scaling of other parts of solver (subspace diag)
- Full code performance

First Principles (Electronic Structure Calculations)

- First Principles: Full quantum mechanical treatment of electrons
- Gives accurate results for Structural and Electronic Properties of Materials, Molecules, Nanostructures
- Computationally very expensive
  (eg. grid of > 1 million points for each electron)
- Density Functional Theory (DFT) Plane Wave Based (Fourier) methods probably largest user of Supercomputer cycles in the world



Ba_2YCl_7:Ce predicted to be a very bright scintillator. Made by experimentalists and found to be one of the brightest known scintillators. Initial Patent Filing taken out for Material Ba_2YCl_7:Ce

CRD Ab initio Method: Density Functional Theory (Kohn 98 Nobel Prize)

Many Body Schrodinger Equation (exponential scaling)

$$\{-\sum_i \frac{1}{2}\nabla_i^2 + \sum_{i,j}\frac{1}{|r_i - r_j|} + \sum_{i,I}\frac{Z}{|r_i - R_I|}\}\Psi(r_1,..r_N) = E\Psi(r_1,..r_N)$$

Kohn Sham Equation (65): The many body ground state problem can be mapped onto a single particle problem with the same electron density and a different effective potential (cubic scaling).

$$\{-\frac{1}{2}\nabla^2 + \int\frac{\rho(r')}{|r - r'|}dr' + \sum_I \frac{Z}{|r - R_I|} + V_{XC}\}\psi_i(r) = E_i\psi_i(r)$$

$$\rho(r) = \sum_i |\psi_i(r)|^2 = |\Psi(r_1,..r_N)|^2$$

Use Local Density Approximation (LDA) for $V_{XC}[\rho(r)]$ (good Si,C)

Plane-wave Pseudopotential Method in DFT (Self-consistent)

$$\boxed{\{-\frac{1}{2}\nabla^2 + \int \frac{\rho(r')}{|r-r'|}dr' + \sum_I \frac{Z}{|r-R_I|} + V_{XC}(\rho(r))\}\psi_j(r) = E_j\psi_j(r)}$$

Solve Kohn-Sham Equations self-consistently for electron wavefunctions within the Local Density Appoximation

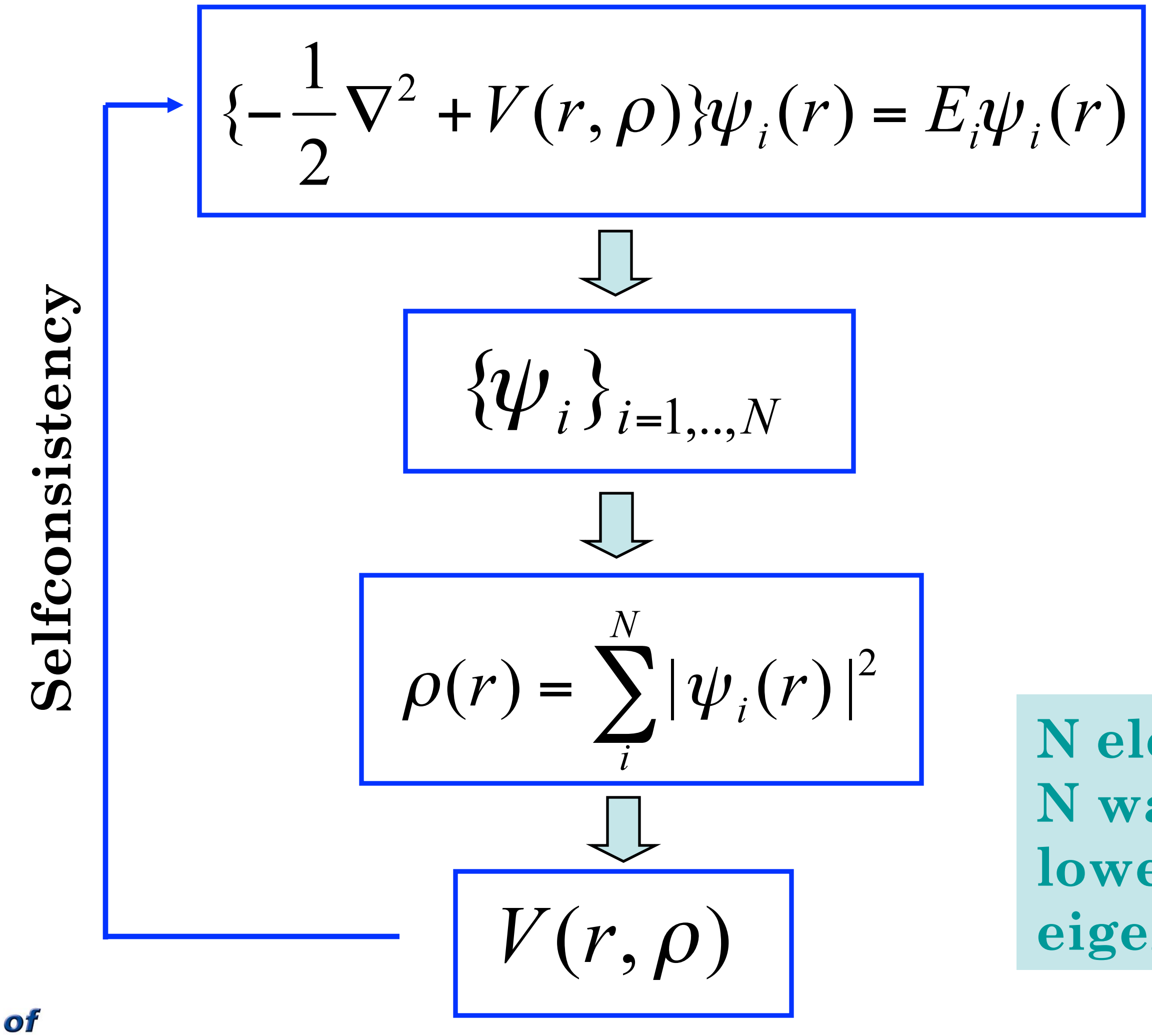
1. **Plane-wave expansion for** $\psi_{j,k}(r) = \sum_g C_g^j(k)e^{i(g+k).r}$

2. **Replace "frozen" core by a pseudopotential**

Different parts of the Hamiltonian calculated in different spaces (Fourier and real) 3d FFT used

Codes: VASP, PARATEC, PeTOT, Abinit, PWSCF, QBox, CASTEP

Self-consistent Calculation



$$\{-\frac{1}{2}\nabla^2 + V(r,\rho)\}\psi_i(r) = E_i\psi_i(r)$$

$$\{\psi_i\}_{i=1,..,N}$$

$$\rho(r) = \sum_i^N |\psi_i(r)|^2$$

$$V(r,\rho)$$

N electrons
N wave functions
lowest N
eigenfunctions

Computational Considerations

$$\{-\frac{1}{2}\nabla^2 + \int\frac{\rho(r')}{|r-r'|}dr' + \sum_I \frac{Z}{|r-R_I|} + V_{XC}(\rho(r))\}\psi_j(r) = E_j\psi_j(r)$$

- Largest DFT type calculations (eg 5,000 Si atoms to calculate dopant levels)
- Matrix size, M = 1.25 million
- Number of required eigenpairs, N = 10,000

- **Matrix never computed explicitly (available through mat-vec product)**
- **Matrix is dense (in Fourier or Real space)**
- **Each SCF step we have good guess for eigenvectors (from previous step)**
- **Want to perform many moderate sized 3d FFTs (512^3 largest systems studied !)**
- **Diagonal KE term dominant, use as preconditioner** $-\frac{1}{2}\nabla^2\psi_i(r) = -\frac{1}{2}g^2\psi_i(r)$

⇨ **Typically use blocked CG based iterative methods (BLAS3)**

Most Costly parts of CG based Solver

$$\{-\frac{1}{2}\nabla^2 + \int \frac{\rho(r')}{|r-r'|}dr' + \sum_I \frac{Z}{|r-R_I|} + V_{XC}(\rho(r))\}\psi_j(r) = E_j\psi_j(r)$$

Computational Task (CG solver)	Scaling
Orthogonalization	MN^2
Subspace (Krylov) diagonalization	N^3
3d FFTs (most communications)	NMlogM
Nonlocal pseudopotential	MN^2 (N^2 real space)

N: number of eigenpairs required $(\psi_j(r), E_j)$ (lowest in spectrum)

M: matrix (Hamiltonian) dimension (M ~ 200N)

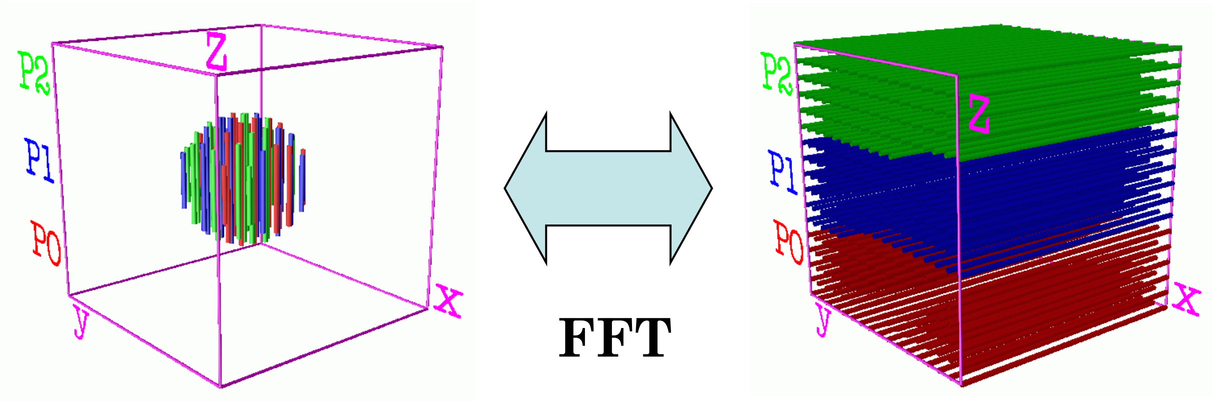
Load Balancing, Parallel Data Layout

- Wavefunctions stored as spheres of points (100-1000s spheres for 100s atoms)
- Data intensive parts (BLAS) proportional to number of Fourier components
- Pseudopotential calculation, Orthogonalization scales as N^3 (atom system)
- FFT part scales as N^2logN

<u>Data distribution: load balancing constraints (Fourier Space):</u>

- each processor should have same number of Fourier coefficients (N^3 calcs.)
- each processor should have complete columns of Fourier coefficients (3d FFT)

$$-\frac{1}{2}\nabla^2\psi_i(r)$$



$$V(r)$$

Give out sets of columns of data to each processor

CRD Parallel 3d FFT on *N*x*N*x*N* (x,y,z) grid

1. Perform 1d FFTs on N^2 x direction columns

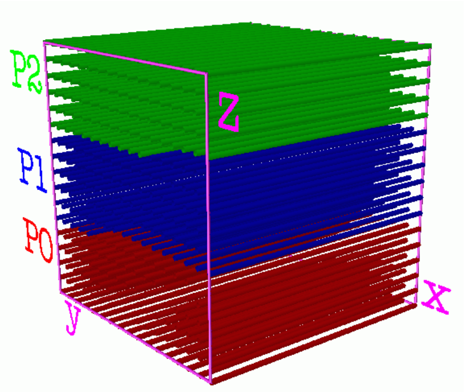
2. Transpose (x,y,z) -> (y,z,x)

3. Perform 1d FFTs on N^2 y direction columns

4. Transpose (y,z,x) -> (z,y,x)

5. Perform 1d FFTs on N^2 z direction columns

6. Transpose (z,y,x) -> (x,y,z) optional

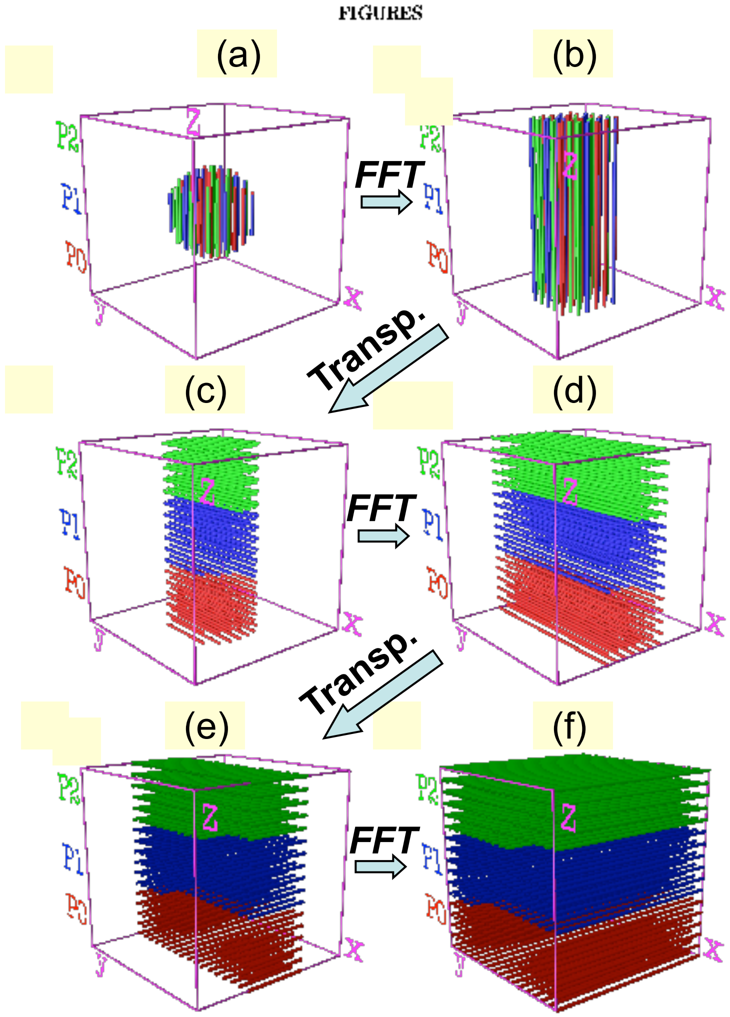


<u>Scaling Issues (bandwidth and latency):</u>

- computations/communications ~ $N^2 N \log N / N^3 = \log N \sim O(10)$

- message size ~ $(\#\text{nproc})^{-2}$ 1d layout $(\#\text{nproc})^{-3/2}$ 2d layout

Specialized 3d FFT for Electronic Structure Codes (Plane Waves/Fourier)



- Works for any grid size on any number of processors
- Only non-zero elements communicated/calculated
- Most communication in global transpose (b) to (c) little communication (d) to (e)
- Much faster than vendor supplied 3d-FFT (no grid size limitations)
- Used in many codes (PARATEC, PETot, ESCAN, GIT code etc.)

Cray XT4 (Franklin, NERSC, LBNL)

- Cray XT4 (NERSC computer center, Lawrence Berkeley Lab.)
- Node: Quad core Opteron 2.3 GHz (peak 9.2 Gflops)
- System: 9,572 compute nodes, 38,288 processor cores
- Interconnect: 3d Torus
- Peak speed: 352 TFlop/sec
- 11th on Top500 list

CRD

Results: 3d-FFT 512^3 grid on Cray XT4

- Strong scaling tests on 512^3 grid forward+reverse 3d FFT
- 512^3 grid corresponds to 1000s atoms in real code, 1000s electrons (grids)
- ~51400 columns in Fourier space for each electron
- Written in Fortran + MPI + FFTW for 1d FFTs
- Versions use MPI_ISENDS and MPI_RECVs/IRECVS or MPI_ALLTOALLV (MPICH2)
- Blocked versions (bl40) perform 40 3d FFTs and aggregate messages (40 times larger)

Procs.	isendrecv	Isendirecv_all	alltoallv	Isendrecv_bl40	alltoallv_bl40
128	0.4139 s	0.3082	0.3605		0.3663
256	0.2730 s	0.1899	0.2123	0.2132	0.1921
512	0.3176 s	0.1725	0.1743	0.1168	0.1004
1024	6.2567 s	0.2499	0.1969	0.1310	0.0558
2048	7.9659 s	0.4469	0.2808	0.2355	0.0370
4096	8.0062 s	0.4726	0.3077	0.3862	0.0312
8192		0.2514	0.2375	0.3263	0.0221
16384			0.1715		0.0136

Very good scaling to 16K procs for alltoallv_bl

Results: Multicore tests for 3d-FFT 512^3 grid on Cray XT4

- Strong scaling tests on 512^3 grid forward+reverse 3d FFT
- 1 to 4 cores per node (each node has Quad core Opteron)
- Memory contention on the node main reason for much slower 4 core performance

Procs. cores	alltoallv_bl (4cores)	alltoallv_bl (2cores)	alltoallv_bl (1core)
128	0.3663	0.2544	0.2120
256	0.1921	0.1301	0.1124
512	0.1004	0.0699	0.0596
1024	0.0558	0.0379	0.0325
2048	0.0370	0.0235	0.0232

Results: Strong Scaling tests for 3d-FFT 512^3 grid on IBM BG/P

- IBM Blue Gene/P system (Intrepid) Argonne National Laboratory
- Node: PowerPC Quad core 450 850 MHz (3.4 GFlops)
- System: 40,960 nodes (163,840 processor cores)
- Peak Speed: 557 Teraflops
- Interconnect, low latency 3D-torus, scalable collective network, fast barrier network
- 7th on top500 list

Procs.	isendrecv	Isendirecv_s	alltoallv	Isendrecv_bl40	alltoallv_bl40
512	0.2413 s		0.1768		
1024	0.1911 s	0.1232	0.0929	0.1377	0.1150
2048	0.9008 s	0.0636	0.0396	0.0843	0.0646
4096	6.5026 s	0.0758	0.0346	0.1611	0.0303
8192	41.494 s	0.0979	0.0342	1.0962	0.0257
16384		0.1175	0.0295	5.1327	0.0124

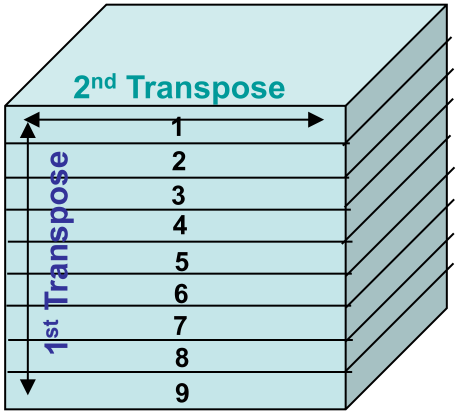
Very good scaling to 16K processors for alltoallv_bl (better than XT4)

Why we don't use a library 3d FFT !

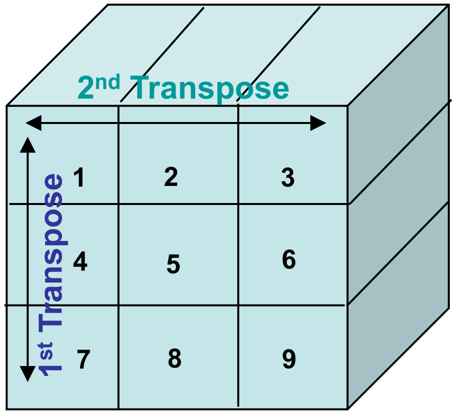
- No 3d FFT libs. that can run any size grid on any number of procs. Grid sizes determined by #atoms (P3DFFT the closest to our needs !)
- Need a complex to complex 3d FFT (P3DFFT is real to complex)
- Would need to transform the data from our load balanced sphere to data layout to use libs. (like an extra transpose)
- No libs. can do blocked 3d FFTs to avoid latency issues
- No libs. Can take advantage of small sphere in Fourier space (we would have to pad system with zeros to full grid size)

Communication costs for transposes (N^3 grid): 1d and 2d processor layout for 3d FFT

1d



2d



1st Transpose:

- Messages: $(\#nproc)^2$ alltoall messages, size: $(N^3)/(\#nproc)^2$

2nd Transpose:

- No communication (#nproc < 512)
- Local limited comms if #nproc > 512

$(\#nproc)^2$ messages, N^3 data transfer (N<512)

1st Transpose:

- Messages: $(\#nproc)^{3/2}$ messages along rows size: $(N^3)/(\#nproc)^{3/2}$

2nd Transpose:

- Messages: $(\#nproc)^{3/2}$ messages along cols. size: $(N^3)/(\#nproc)^{3/2}$

$2(\#nproc)^{3/2}$ messages, $2N^3$ data transfer

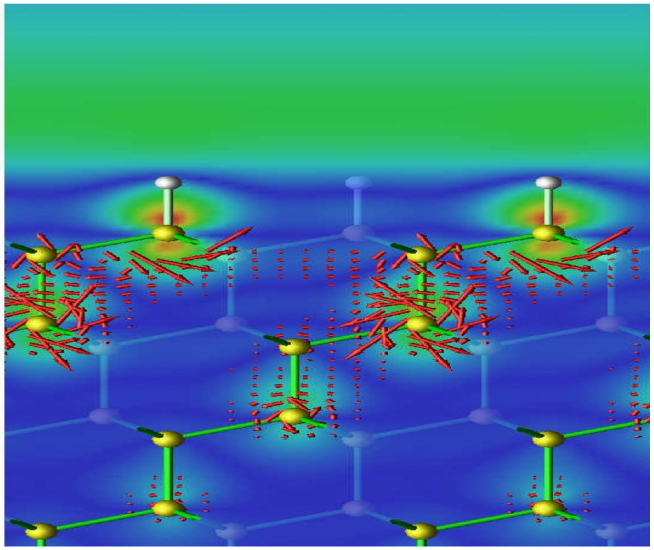
Comparison to P3DFFT and 3d FFTW on Cray XT4

Strong scaling tests on 512^3 grid forward+reverse 3d FFT
Time for P3DFFT real to complex doubled, time in brackets is for real to complex

Procs.	alltoallv_bl40	P3DFFT [1d proc. layout]	P3DFFT [2d proc layout]	3d FFTW
128	0.3663 s	0.4988 (0.2494) [1x128]	1.0498 (0.5249) [8x16]	1.1275
256	0.1921 s	0.3228 (0.1614) [1x256]	0.5450 (0.2725) [16x16]	0.6235
512	0.1004 s	0.2938 (0.1469) [1x512]	0.2824 (0.1412) [16x32]	1.4063
1024	0.0558 s	0.3050 (0.1525) [2x512]	0.1236 (0.0618) [32x32]	
2048	0.0370 s	0.2370 (0.1185) [4x512]	0.0766 (0.0383) [32x64]	
4096	0.0312 s	0.2154 (0.1077) [8x512]	0.0698 (0.0349) [64x64]	
8192	0.0221 s	0.1659 (0.0829) [16x512]	0.0874 (0.0437) [64x128]	
16384	0.0136 s		0.0958 (0.0479) [128x128]	

Absolute performance and scaling is much better for our 3d FFTs
(P3DFFT does not scale past 2K processors)

CRD PARATEC (PARAllel Total Energy Code)



- PARATEC performs first-principles quantum mechanical total energy calculation using pseudopotentials & plane wave basis set
- Written in F90 and MPI
- Designed to run on large parallel machines IBM SP etc. but also runs on PCs

- PARATEC uses all-band CG approach to obtain wavefunctions of electrons (blocks comms. Specialized 3dffts)
- Generally obtains high percentage of peak on different platforms (uses BLAS3 and 1d FFT libs)
- Developed with Louie and Cohen’s groups (UCB, LBNL)

Overall ratio calcs./communications ~ N (not logN)

PARATEC: Performance

Problem	Proc	Bassi NERSC (IBM Power5)		Jaquard NERSC (Opteron)		Thunder (Itanium2)		Franklin NERSC (Cray XT4)		NEC ES (SX6)		IBM BG/L	
		Gflops/ Proc	% peak	Gflops/ Proc	% peak	Gflops/ Proc	% peak	Gflops/ Proc	% peak	Gflops/ Proc	% peak	Gflops/ Proc	% peak
488 Atom CdSe Quantu m Dot	128	5.49	72%			2.8	51%			5.1	64%		
	256	5.52	73%	1.98	45%	2.6	47%	3.36	65%	5.0	62%	1.21	43%
	512	5.13	67%	0.95	21%	2.4	44%	3.15	61%	4.4	55%	1.00	35%
	1024	3.74	49%			1.8	32%	2.93	56%	3.6	46%		
	2048							2.37	46%	2.7	35%		

- Grid size 252^3
- All architectures generally achieve high performance due to computational intensity of code (BLAS3, FFT)
- ES achieves highest overall performance : 5.5Tflop/s on 2048 procs (5.3 Tflops on XT4 on 2048 procs in single proc. node mode)
- FFT used for benchmark for NERSC procurements (run on up to 18K procs on Cray XT4, weak scaling)
- Vectorisation directives and multiple 1d FFTs required for NEC SX6

Office of Science U.S. DEPARTMENT OF ENERGY

Developed with Louie and Cohen’s groups (UCB, LBNL), also work with L. Oliker, J Carter

PARATEC: Performance (new code)

Problem	Proc	Franklin NERSC (Cray XT4)	
		Gflops/ Proc	speedup
488 Atom CdSe Quantum Dot	128	304.7s	1.0 (1)
	256	177.3s	1.72 (2)
	512	84.33s	3.61 (4)
	1024	43.25s	7.05 (8)
	2048	25.93s	11.75 (16)
	4096	20.09s	15.16 (32)

- Grid size 252^3 (larger system 1000 atom being run will give better scaling on other parts of code)
- Need to recode many other parts of code so memory etc. scales better

Other plane wave DFT code:

- QBox (also CPMD) get higher levels of scaling via 3 level parallelism:
- QBox Gordon Bell SC06: 64K nodes on BG/L (207 TFlops) 1000 atoms metal (larger than our system)
- 64k = (8 k points) x (16 bands) x (512 for 3d FFT)

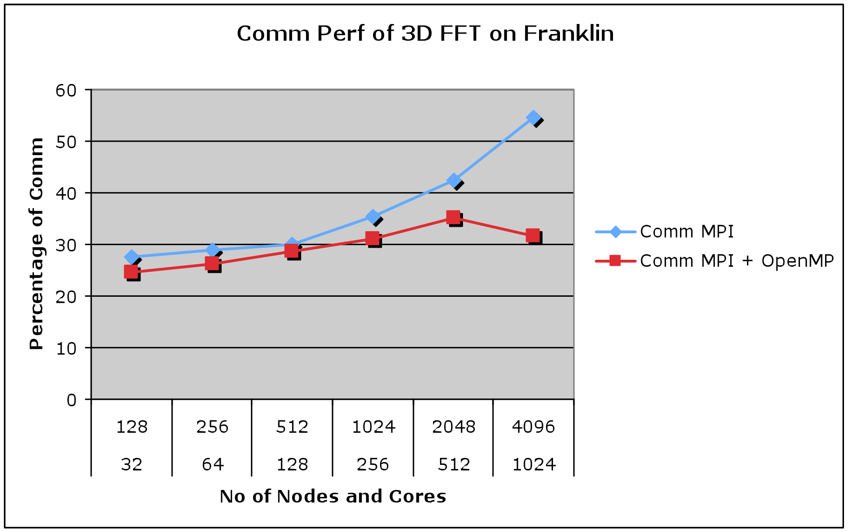
CRD

3d FFT Mixed OpenMP/MPI version

Motivation: One MPI process per node allows us to send fewer larger messages (n^2_{nodes} vs. $n^2_{tot\#cores}$)

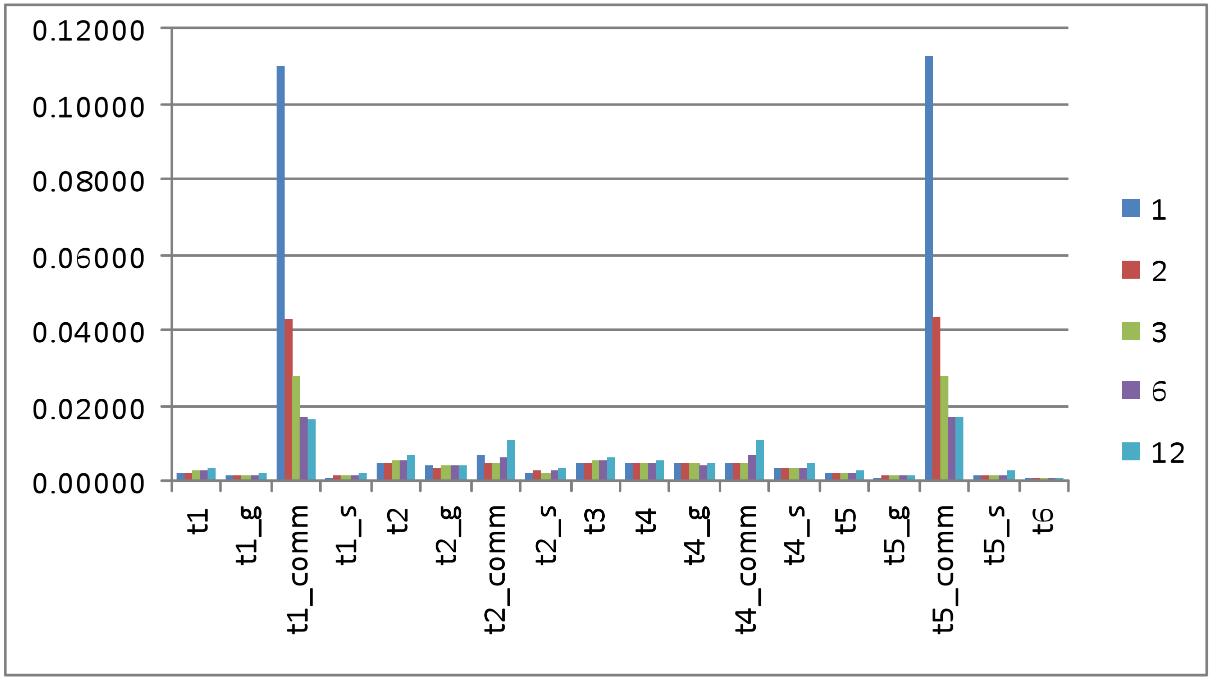
Three computationally distinct parts

1. **1d FFTs Parallelizes well with OpenMP (similar performance to pure MPI version)**
2. **Gather/Scatter operations used before and after communications to perform transposes OpenMP version slower than pure MPI (small work load for each thread)**
3. **MPI alltoall communication step** (large gain from fewer, larger messages)



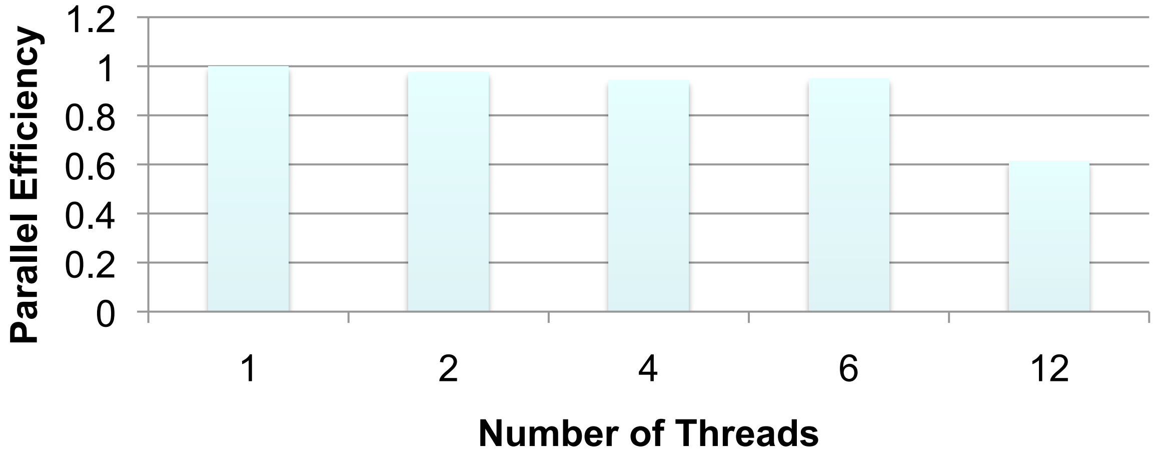
3d FFT Mixed OpenMP/MPI version (Jaguar)

Packed 576 cores 1-12 threads (Forward and Reverse FFT)



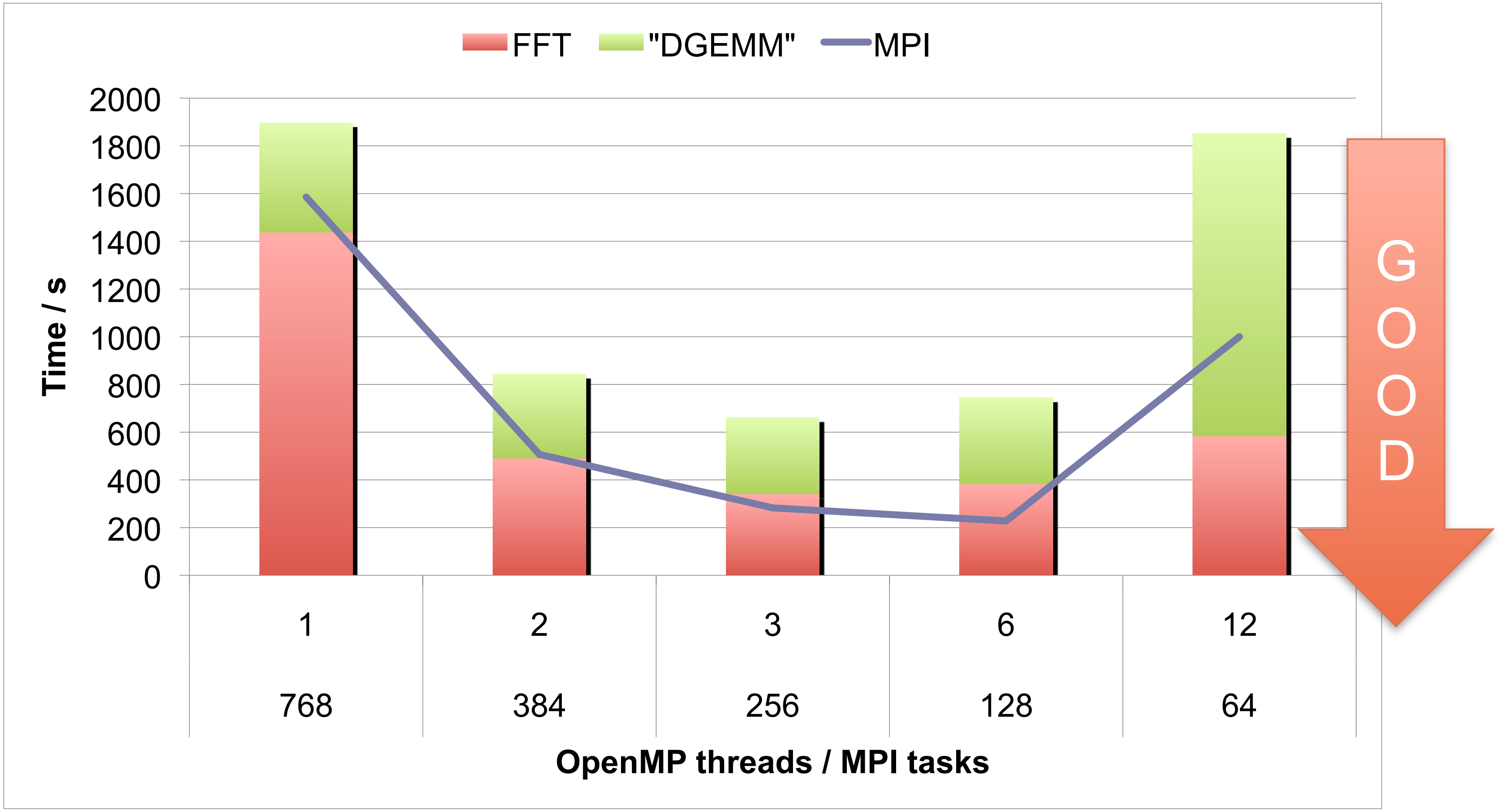
OpenMP/MPI in PARATEC

- PARATEC 30-40% ZGEMM very amenable to threading

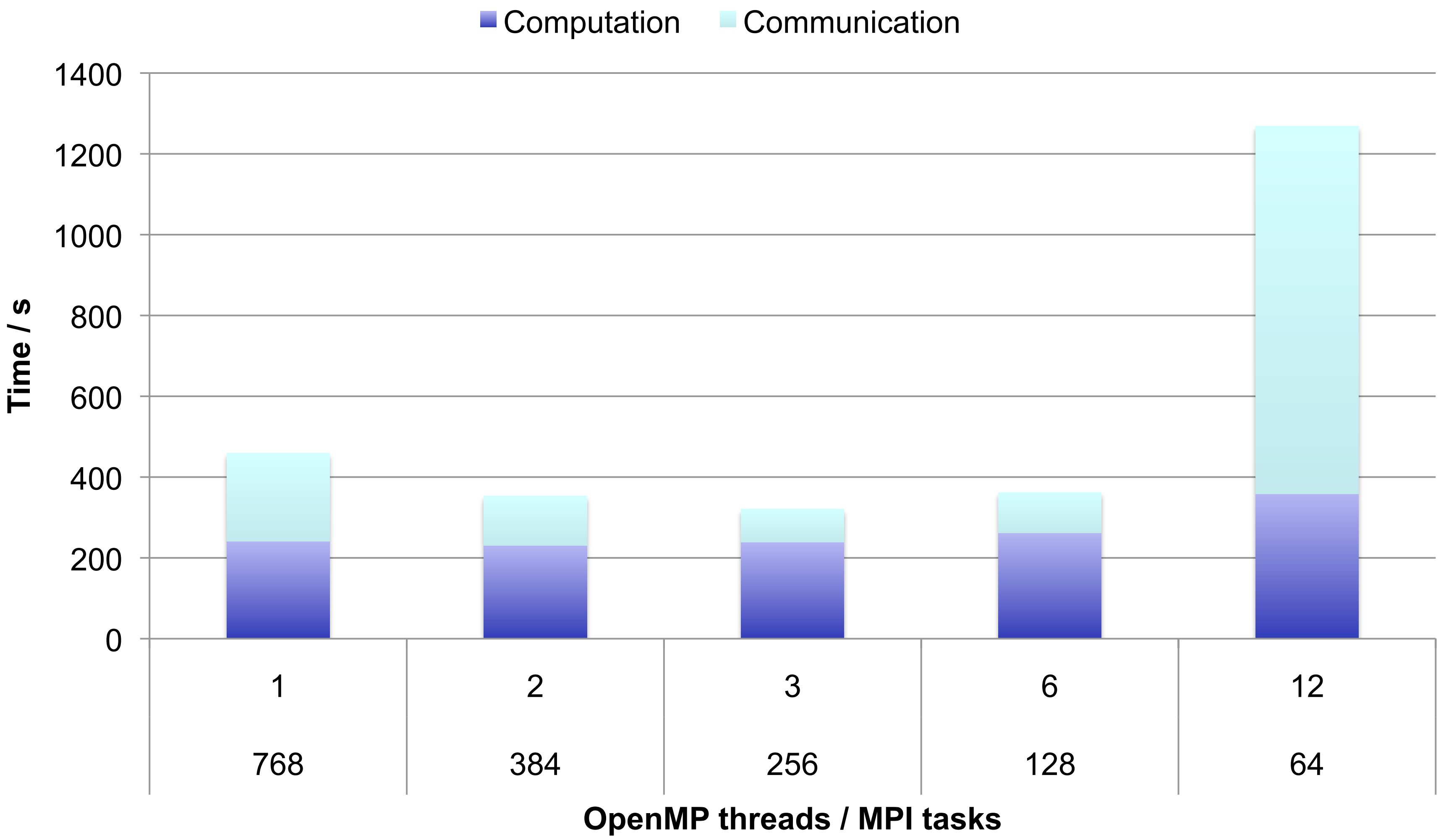


- Can aggregate messages in other parts of code

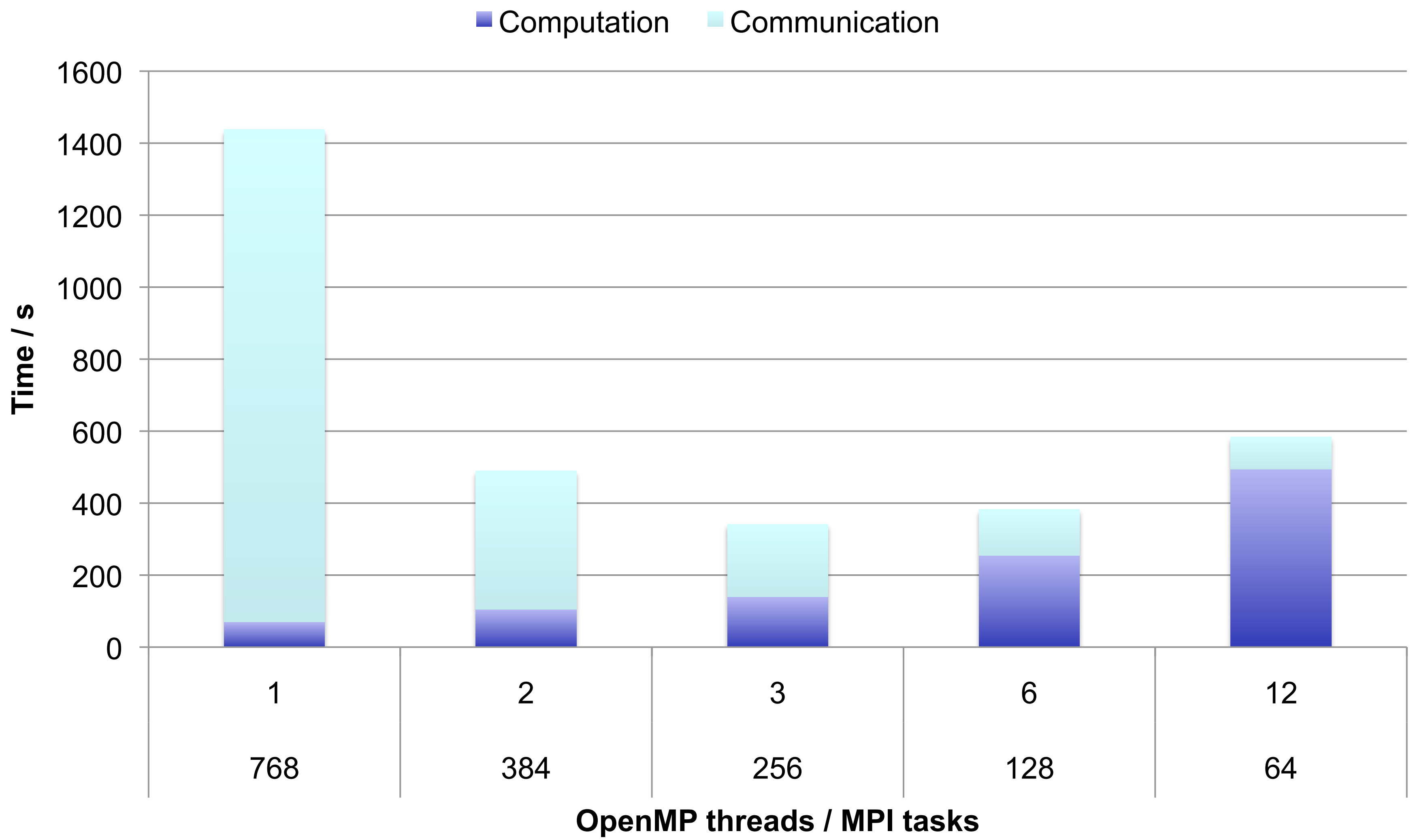
Paratec MPI+OpenMP Performance (Jaguar)



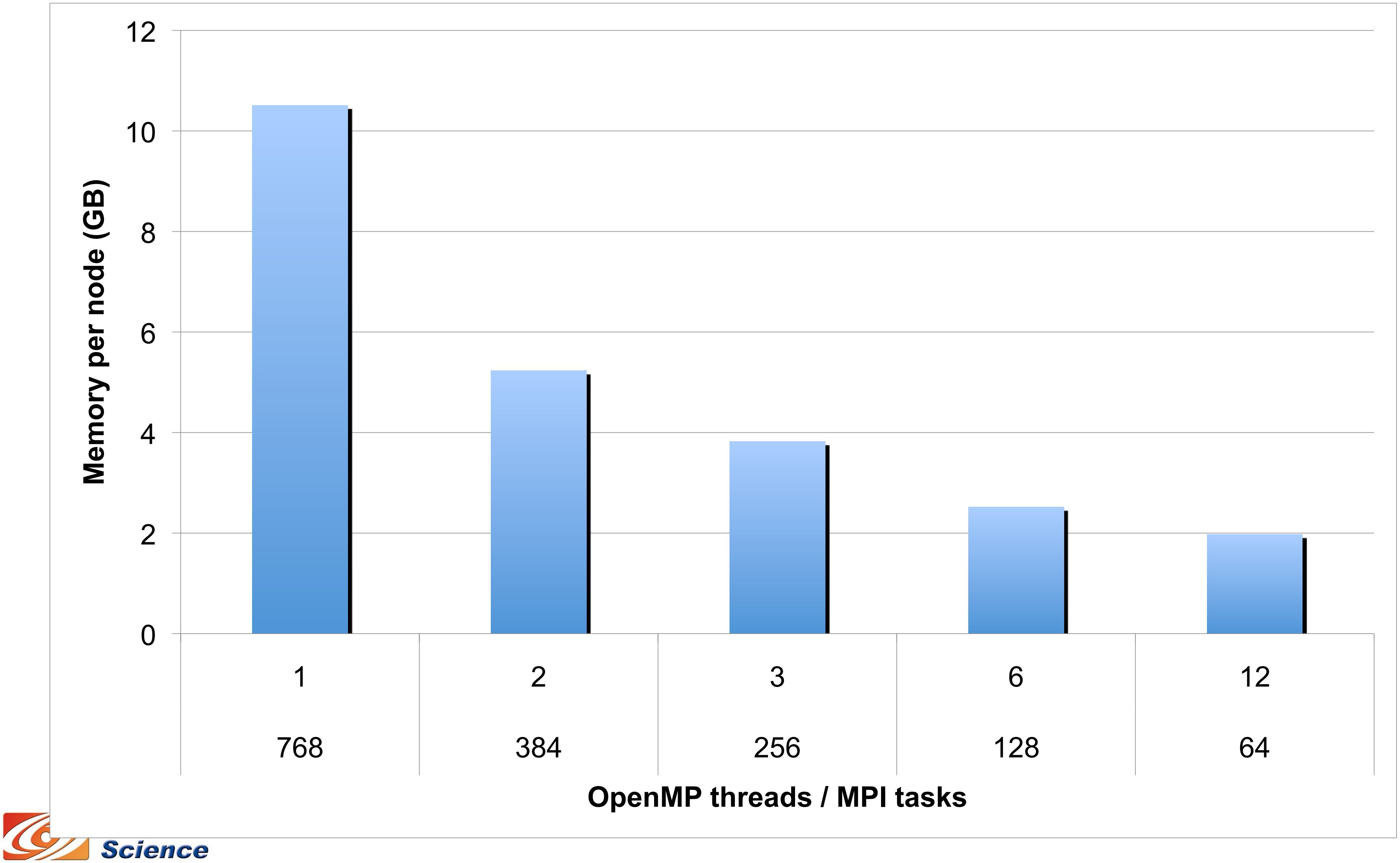
Non FFT part of code “ZGEMM”



FFT Breakdown



PARATEC - Memory Usage



Subspace Diagonalization

Computational Task (CG solver)	Scaling
Orthogonalization	MN^2
Subspace (Krylov) diagonalization	N^3
3d FFTs (most communications)	NMlogM
Nonlocal pseudopotential	MN^2 (N^2 real space)

Diagonalization Problem: matrix size may be of the order of the number of processors

Solution: run on the number of procs that corresponds to: min. block size of 32-64 and as close as possible to a square processor grid to get best possible speedup for scalapack

VASP code

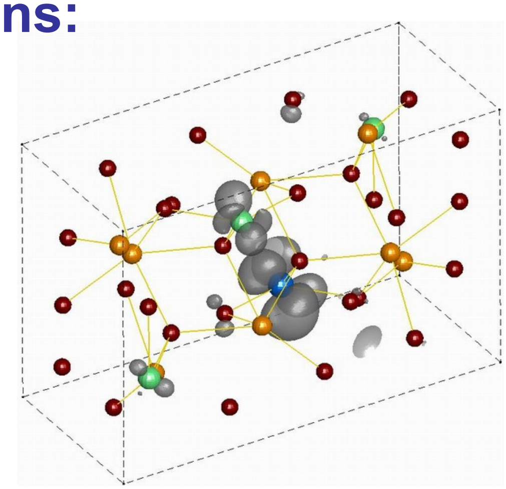
- Supports many different methods and features (Ultrasoft pseudopotentials, PAW, HF, Hybrid functionals)
- Supports plane wave coeffs. (g vector) and band parallelization
- Default minimization is band by band CG (cannot aggregate messages in FFT, cannot use band parallelization, cannot use BLAS3)
- Residual minimization supports band parallelization (and aggregation in FFT, P. Kent)

CRD

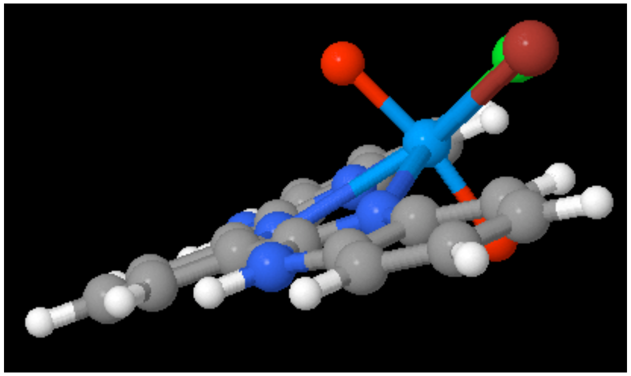
Summary and Future Directions

- Fourier electronic structure (3d FFTs) can scale to 16 K processor regime (not limiting factor in scaling !) also allow Qbox, VASP etc. to scale to higher number of procs.
- Future directions: threads on node (for 1d FFTs), overlap calcs/comms etc.

Applications:



New gamma ray detector materials



New ligands for nuclear waste separation

Office of Science
U.S. DEPARTMENT OF ENERGY